FO 103006 RR (3)

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(11) Publication number : **0 453 355 A2**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **91400976.6**

(51) Int. Cl.^{5} : **H04L 12/56**

(22) Date of filing : **11.04.91**

(30) Priority : **13.04.90 US 509605**

(43) Date of publication of application :
23.10.91 Bulletin 91/43

(84) Designated Contracting States :
DK FR GB IT

(71) Applicant : **DIGITAL EQUIPMENT CORPORATION**
146 Main Street
Maynard, Massachusetts 01745 (US)

(72) Inventor : **Goldstein, Fred Richard**
279 Appleton Street
Arlington, Massachusetts 02174 (US)
Inventor : **Callon, Ross**
5 Hilda Road
Bedford, Massachusetts 01730 (US)

(74) Representative : **Mongrédien, André et al**
c/o SOCIETE DE PROTECTION DES INVENTIONS 25, rue de Ponthieu
F-75008 Paris (FR)

(54) Congestion avoidance in high-speed network carrying bursty traffic.

(57) A data communication network subject to bursty traffic employs a bandwidth allocation scheme to avoid congestion. When a source node has a burst of traffic to send, it first sends a bandwidth request message through the network from source to destination. At each intermediate node, this bandwidth request is examined and the node determines how much of the requested traffic level it will be able to support at a time in the future of one round-trip interval hence, and this node either grants the request or marks down the request to a level that it can support, then passes it on. When the request reaches the destination, it is returned along the same path to the source, and the source then employs the marked-down allocation to select the rate used to send the burst of data. The allocation for this source node remains in effect for a limited time, depending upon the amount of data to be sent in the burst, then returns to a "residual" level.

EP 0 453 355 A2

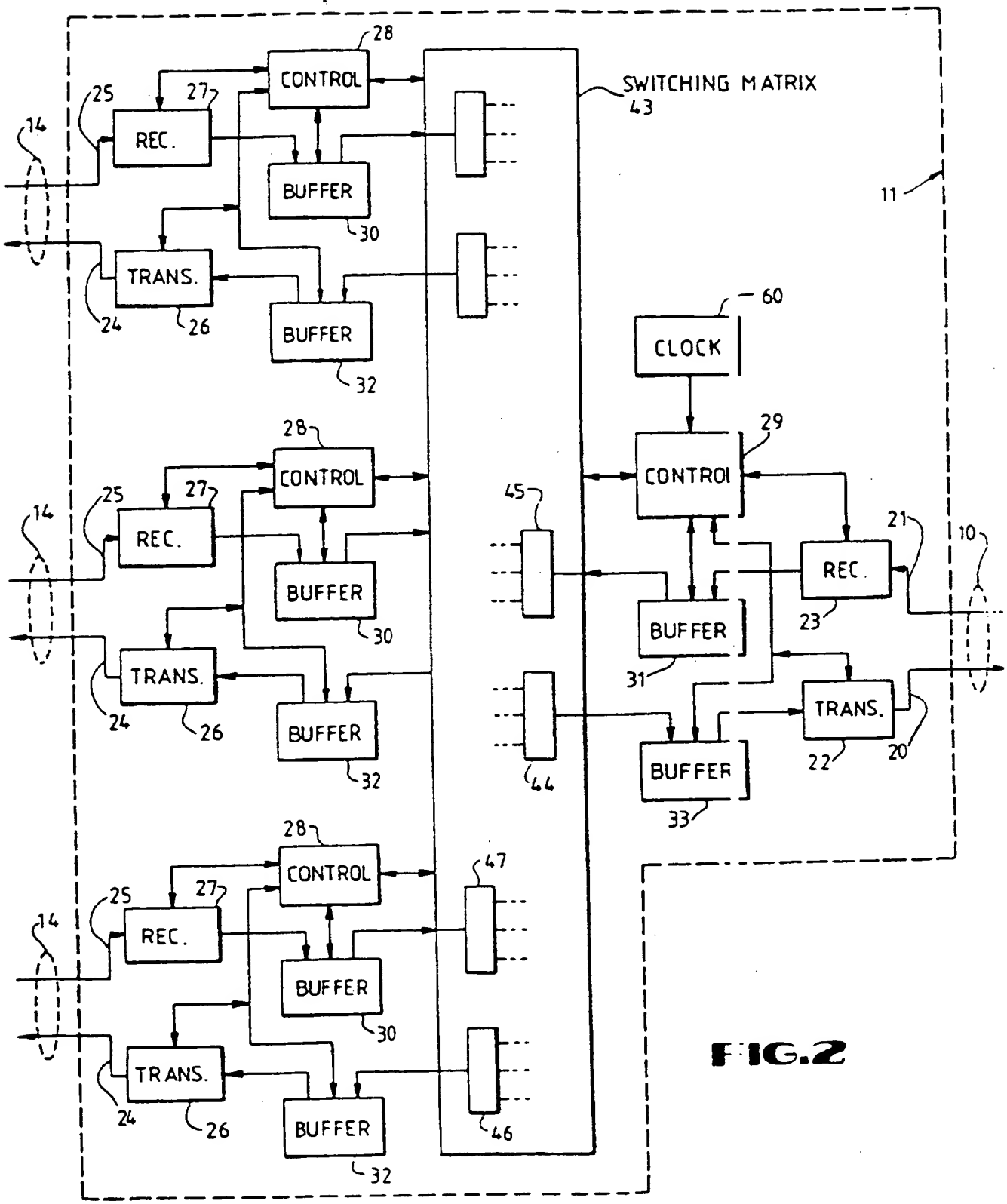


FIG.2

BACKGROUND OF THE INVENTION

This invention generally relates to data communication networks, and more particularly to congestion avoidance in networks having links exhibiting long propagation delay and having bursty data flow.

Broadband ISDN (integrated services digital network) systems are prone to severe buffer overflow problems at intermediate nodes. Data is thus lost and must be retransmitted, reducing the reliability and capacity of the system. These losses are referred to as congestion loss, but this is not the result of an under-engineered network. Instead, this type of congestion is an inevitable result of the bursty nature of data in asynchronous (packetized) information transfer.

In narrowband packet networks, feedback control mechanisms are able to manage the traffic load so that buffer overflow can be mostly avoided, or at least controllable. For example, networks based upon the so-called X.25 protocol provide two levels of flow control; one controls all traffic across the physical link; and another layer controls traffic across one virtual circuit. Each layer's protocol provides both a window-based flow manager and a stop/go control mechanism. That is, a node having data to send is allocated a window of fixed time, per unit time, and, in addition, the node can be shut off for a time period when capacity is all allocated.

Connectionless networks, such as those using certain DECnet and DDN Internet Protocol, do not have positive controls as in X.25, but still provide positive feedback mechanisms. For example, "implicit" feedback mechanisms focus on sequence numbers in the packets; if a packet is dropped (as may be inferred from a gap in acknowledged sequence numbers) it may be determined that congestion is severe and so the sender drastically reduces its sending rate (as by reducing the window size). Or, "explicit" mechanisms provide warning of incipient congestion, so that senders can usually reduce their rate before any packets are lost; thus there is still feedback, but the data terminals are more responsible for responding to it.

Broadband asynchronous transfer mode (ATM) networks often have links that span large distances, thousands of miles in many cases. Here the propagation delay is too long to allow feedback to be effective. The delay from the time a packet is sent to the time it is received at the destination is much longer than the time during which congestion can cause buffers to fill in intermediate nodes, so data is lost. By the time the loss is recognized and a feedback signal sent back to the sender, it is too late to alter the sending rate or otherwise change the input to prevent congestion.

It is not sufficient to use any of the common feedback schemes, including credit managers, windows, etc., across long-delay ATM networks. While some of these techniques are quite appropriate for short-haul ATM applications, they lose effectiveness when the buffer fill time falls well below the propagation delay in the link. The exact point at which feedback delay becomes unacceptable depends upon the degree of burstiness of the traffic; if the bulk of traffic is constant, then a somewhat longer delay can be tolerated before loss occurs. Highly bursty traffic is more sensitive.

In a typical network, therefore, at least two types of bandwidth allocation are needed. A simple credit-based buffer allocation scheme is likely to be quite adequate for certain applications - those links that have short propagation delay. The receiving end of each link monitors the buffers available to each virtual path and/or virtual channel and grants credits to the sending end. This may be somewhat cumbersome when the number of virtual paths is quite high, but in practice credits are allocated back to each virtual path or channel based upon availability of buffers for the links going out of the node. Some amount of receive buffer may be useful in order to permit a node to accept all traffic arriving at one incoming link when there is a disparate buffer fill situation at its outgoing links. Nonetheless, this transmission discipline is simply a form of conventional hop-by-hop management, and is not dis-similar from what is found on conventional connection-oriented packet networks. These links may not need a more complex scheme such as is described below. A more complex discipline is only required when the dimensions of the network cause propagation delays to become longer than allowable feedback times.

There are two causes of congestion loss, funnelling and mismatch. A packet-switched network can lose its protocol data units (cells, frames or packets) when the arrival rate at any given point exceeds the departure rate for a long enough period of time that a buffer overflows. This can occur for either of two separate and identifiable reasons. Funnels occur when several different paths converge on a single buffer, and traffic bursts arrive closely spaced in time, such that overflow occurs. Funneling is generally transient. Mismatch occurs when sustained demand for a given facility exceeds its capacity; for example, when a high-speed link meets a lower-speed link, or when an additional virtual circuit is created over a busy facility. A congestion management scheme must be able to handle both mismatch and funneling. However, deferent techniques tend to be more effective for one or the other. Admission control policies, coupled with stringent network-wide resource allocation and a minimum of oversubscription, can minimize mismatch. Connectionless networks require feedback to control mismatch loss, as they rarely if ever provide rate-based control. Traditional packet networks are usually rather tolerant of funneling loss. An occasional dropped packet can be recovered. ATM networks, however, may use protocols that are prone to loss multiplication; a single dropped cell can corrupt an

entire packet, if frame-based recovery is used. Thus, funneling effects are far more severe in an ATM context, and are hardest to solve.

It has been suggested that by limiting the rate at which users are allowed to send data into an ATM network (i.e., access control) such that the total bandwidth of all channels does not exceed the cross-sectional size of any trunk facility, then congestion will not occur. This is not true, however, when the traffic is bursty. While bursts in an ATM network may be individually bounded in size and rate, a probability exists that at any given time the amount of traffic arriving for any given buffer will exceed the capacity of that buffer, even if the average is not excessive.

One cause of funneling effect is most likely when many small virtual channels are provided. If the number of virtual channels exceeds the number of cells in a buffer, then it is statistically possible that all of them send their cells at such times that the bursts arrive at a given buffer close enough in time that the buffer overflows. The total "event horizon" within an ATM network is no greater than the longest round-trip delay including buffer times. Thus, even circuits with a "reserved" throughput class (enforced at the access) of, say, 64- kbps, who are thus allowed to issue one cell (with a 48-octet payload) every 6-ms., can send those cells anywhere within the 6-ms. window, and indeed more likely will be allowed to accumulate credits so that larger bursts may be sent with less frequency.

A second cause of funneling loss is the simultaneous arrival of multiple large bursts of traffic at a common point. High-speed data traffic, such as occurs on local area networks and which may migrate to ATM or other wide are: networks, is characterized by bursts of data at or near the data rate of the physical facility, and consisting of more information than might fit into a typical ATM network buffer. Even two such bursts of data converging onto one facility may result in funneling loss, as the buffer is smaller than the bursts and data is arriving more rapidly than it is exiting the buffer.

It can therefore be shown that no access control scheme can positively prevent buffer overflow. If ATM networks are to use loss- sensitive protocols, then a different mechanism is required to prevent cell loss. Such a mechanism must actively counter the bursty nature of ATM traffic, to reduce peak buffer occupancy and thus the chance of overflow. Continuous bit rate services are in this respect little different from variable bit rate services, because typical variable bit rate user variations in rate occur over a time period that is quite long, compared to sub-millisecond buffer fill times.

Variable bit rate is thus handled by treating it as a special case of continuous bit rate, in which the bit rate is changed on occasion. Most bursty data can tolerate delays in the 100-ms. range; if realloca- tion of bandwidth takes this long, applications and users will typically not notice. The invention is thus capable of operating over circuit- switched networks with high switching speed, as well as over ATM networks.

Circuit-switched digital networks (including narrow bandwidth integrated services digital networks) typically make use of synchronous time division multiplexing to allocate bandwidth on trunk facilities; in this method, individual channels are separated by their position within a stream. Some bandwidth is required for framing purposes, but individual channels have no overhead of their own. Asynchronous transfer mode (ATM) is, in effect, "asynchronous time division multiplexing", where individual channels are identified by label, instead of by position within the stream. Asynchronous transfer mode is thus more akin to packet mode in operation, although it operates below the data link layer and does not provide the same services as narrowband packet networks.

Because the bursty nature of ATM will necessarily result in buffer overflow in heavily-loaded (but not necessarily oversubscribed) networks with long-delay physical facilities, buffer management (i.e., congestion avoidance) requires a redefinition of the problem, as addressed by the present invention.

SUMMARY OF THE INVENTION

A technique is described herein that is compatible with asynchronous transfer mode, uses ATM-type labeled cells and provides similar services, but is not completely asynchronous; this technique is applicable to variable bit rate applications.

The invention herein described may be use d in various types of networks; one network which may use the features of the invention employs a "plesiochronous" transfer mode (PTM) (plesiochronous=near synchronous) The plesiochronous transfer mode provides services like asynchronous transfer mode (ATM) but with much lower (albeit non-zero) probability of cell loss, and is intended for use on long-delay links within ATM networks. PTM uses cells like ATM, but prevents buffer overflows by pre-allocating bandwidth. Like synchronous time division multiplexing, time slots are used, but here the slots are one cell wide (53-octets) and labeled with cell headers. Slotting is thus used solely as a buffer management and congestion control mechanism. The operation of a plesiochronous transfer mode link is similar to that of a phase locked loop, with each ptm link operating with a fixed frequency (periodicity). Within each PTM link, a fixed number of slots, each carrying one ATM cell, is provided. Individual virtual channels are assigned slots, based upon the required bandwidth. Because operation is plesiochronous (near-synchronous) and not synchronous, buffering is still required.

The invention in its broad form resides in a method of congestion avoidance in a communications

network having multiple nodes, comprising the steps of: a) sending from a first of said nodes to a first intermediate one of said nodes a request for an allocation of bandwidth for a transmission of a quantity of data from said first node to a second one of said nodes via said first intermediate node and a second intermediate node; characterized by: b) comparing said request in said first intermediate node with capacity at said first intermediate node to meet said request, and generating a modified request to reduce said allocation if necessary; c) sending said modified request from said first intermediate node to said second intermediate node; d) comparing said modified request in said second intermediate node with capacity at said second intermediate node to meet said modified request, and generating a second modified request to reduce said allocation if necessary; e) sending said second modified request back to said first node; and f) transmitting said quantity of data from said first node to said second node using the bandwidth specified in said second modified request.

The invention also consists in a communications network having multiple nodes, comprising: a) a transmitter sending from a first of said nodes to a first intermediate one of said nodes a request for an allocation of bandwidth for a data transmission from said first node to a second one of said nodes via said first intermediate node and a second intermediate node; b) means for comparing said request in said first intermediate node with capacity at said first intermediate node to meet said request, and for generating a modified request to reduce said allocation if necessary, said modified request being sent from said first intermediate node to said second intermediate node; c) means for comparing said modified request in said second intermediate node with capacity at said second intermediate node to meet said request, and for generating a second modified request to reduce said allocation if necessary, said second modified request being sent back to said first node; d) said transmitter sending data from said first node to said second node using the bandwidth specified in said second modified request.

In accordance with one embodiment of this invention, a data communication network of the type subject to bursty traffic and having long-delay links employs an asynchronous transfer mode in which small, fixed-length blocks of information (cells) are transferred at very high speed. The network employs a bandwidth allocation scheme to avoid congestion. When a source node has a burst of traffic to send, it first sends a bandwidth request message through the network from source to destination. At each intermediate node, this bandwidth request is examined and the node determines how much of the requested traffic level it will be able to support by reservation at a time in the future of one round-trip interval hence, and this node either grants the request or marks down the request to a level that it can support, then passes it on. When the request reaches the destination, it is returned along the same path to the source, and the source then employs the marked-down allocation to select the rate used to send the burst of data. The allocation for this source node remains in effect for a limited time, depending upon the amount of data to be sent in the burst, then returns to a "residual" level.

BRIEF DESCRIPTION OF THE DRAWINGS

The novel features believed characteristic of the invention are set forth in the appended claims. The invention itself, however, as well as other features and advantages thereof, will be best understood by reference to the detailed exemplary description of specific embodiments which follows, when read in conjunction with the accompanying drawings, wherein:

Figure 1 is an electrical diagram in block form of a communica- tions network in which one embodiment of the invention may be utilized;

Figure 2 is an electrical diagram in block form of one of the nodes in the network of Figure 1;

Figure 3 is a diagram of the format of a packet which may be employed in some links of the network of Figure 1;

Figure 4 is a timing diagram of the format of a frame or synchronized loop which may be employed in a long-delay link of the network of Figure 1, according to the PTM technique;

Figure 5 is a diagram similar to Figure 1 of a part of a network having two long-delay links; and

Figure 6 is a timing diagram of a series of the frames of Figure 4 transmitted by one node of Figure 5 and received by another node;

Figure 7 is a timing diagram of an allocation request message used in the network of Figure 1 or 5, according to a feature of the invention; and

Figure 8 is a diagram of message traffic as a function of time in a system of Figures 1 or 5.

DETAILED DESCRIPTION OF SPECIFIC EMBODIMENT

Referring to Figure 1, a communications network is illustrated having a communications link 10 between a pair of nodes 11 and 12. The link 10 is a trunk facility, where the link 10 must carry traffic between a large number of nodes 13 connected to the node 11 by links 14 and a large number of nodes 15 connected to the node 12 by links 16. In turn, the nodes 13 and 15 are connected to other nodes 17 and 18. Thus, because of the fan-in of a number of nodes to the nodes 11 and 12, the link 10 must have much greater capability than the link 19 between nodes 13 and 17, for example. The link 10 may be a satellite link, or fibre optic link, which may span hundreds or thousands of miles, and would have a speed of perhaps 600-

Mbit/sec or more, i.e., broadband. The links between nodes 11, 13 and 17 (or 12, 15 and 18), however, may be broadband facilities operating at perhaps 150- or 600-Mbps, or other speeds, or may be narrowband facilities, i.e., 64000-bps.

The network of Figure 1 may carry both voice and data, with the voice messages being in digital format, so the network may be of the ISDN or integrated services digital network type. Data channels are characterized by sporadic bursts of information, and by one-way or two- way transmissions at any one time, over a wide range of average bandwidth. Thus, a large number of channels may be funneled through a broadband trunk facility such as the link 10 of Figure 1. Voice channels send data at a constant bit rate.

While the link 10 is broadband, there is nevertheless a limit on the amount of traffic it can handle, so some type of control on access by the nodes must be imposed. A simple method of control would be to determine the number of nodes in the entire network at a given time and divide the available bandwidth equally (introducing priorities if needed). However, this method would not take advantage of the statistical gain inherent in a shared data path; many or most of the nodes would not be using their allotted bandwidth at a given moment, so much of the network capacity would be sitting idle while some nodes needing heavy message traffic would be delayed in their operation. Oversubscribing to take advantage of the statistical gain, however, results in congestion at times of heavy traffic, with loss of data due to buffers filling up at intermediate nodes.

It has been found advantageous to employ an asynchronous transfer mode (ATM) in networks of the type seen in Figure 1. That is, the data flow in the link 10 is buffered at nodes 11 and 12 and transmissions are not necessarily synchronized with the originating node 17 or 18. A block of the 64000-bps data from a node 17 may be buffered at node 11 and sent to node 12 at 150- or 600-Mbps (depending upon the network construction) when it is convenient for the node 11 to do so (unrelated to the clock used by the node 17 to send data to the node 13, or by the node 13 to send data to node 11), yet the data rate of the trunk facility is so high that no delay is perceptible at the originating node 17.

The time delay inherent in the link 10 may be too long to allow any meaningful feedback from the node 11 to the node 12 about the receive conditions. If the link 10 is a satellite link, this delay would be several hundred milliseconds. If the link 10 is fibre optic, then the speed of transmission is about 200km/msec., so a link across the continent creates a delay of many milliseconds, which is still too long for feedback to be effective. If the node 11 had a buffer for each of its egress ports 14 of a size of 150 cells, for example, where each cell is 424 bits (53 octets) long, and the link 10 had a payload of 150- Mbps, then if a two-to-one mismatch occurred at one buffer (one of the ports to links 14) this buffer would go from empty to full in about (150 X 424)/(1.5 X 105) = 0.424 milliseconds. It is thus seen that a feedback scheme for control of traffic flow would be inadequate when propagation time delays measured in milliseconds are prevalent. A fundamental rule of control theory is that feedback, to be effective, must be received quickly enough so that response can be timely; delayed too much, feedback will not have the desired effect. In broadband communications networks using asynchronous transfer mode, as discussed with reference to Figure 1, operating over wide area topology, the event being controlled (a buffer filling up at a destination or at some intermediate point) can occur before the feedback (traveling at the speed of light, or at the signal speed in the medium in question) can reach the source node (the point to be controlled).

Thus, due to the vast difference in propagation delays among the links in Figure 1, the plesiochronous transfer mode as herein described might be used for the link 10 which is long-delay, while credit-based ATM (i.e., a shutter or "loose window") is used for the short-delay links 14 or 16, for example.

Referring to Figure 2, a typical construction of one of the nodes 11 or 12 is shown, and the nodes 13 and 15 may be similarly constructed. Although the node 11 of Figures 1 or 2 is illustrated to have three links 14 and one link 10 as the ingress and egress ports, it is understood that the node 11 may have many more ports than four. The link 10 has a transmit line 20 and a separate receive line 21, and these are connected to a transmitter 22 and a receiver 23, respectively, in the node 11. Similarly, each one of the links 14 has a transmit line 24 and a separate receive line 25, and again each of these is connected to a transmitter 26 or receiver 27, respectively. Although it is not necessarily the case, the link 10 in this example is of broader bandwidth (higher rate of transmission) than the links 14, so traffic is funneled into link 10 from several links 14. The function of the receivers 23 or 27 is to detect and demodulate the signal on the line 21 or 25, recover the clock, convert the serial data on the line to parallel data for loading into the receive buffers. The function of the transmitters 22 or 26 is to move data from a buffer in parallel format, convert the data from parallel to serial, modulate a carrier with the serial data, and send the data signal out on the transmit line 20 or 24. Each one of the ports for links 14 is operated by a controller 28, and the port for the link 10 is operated by a controller 29. These controllers are usually processors executing code stored in local memory, or may be state machines, or similar logic. A receive buffer 30 is provided for incoming data from the line 25 for each of the links 14, and likewise incoming data on the line 21 of the link 10 is buffered in a buffer 31. A transmit buffer 32 may also be provided for outgoing data on lines 24 for each link 14, as well

as a transmit buffer 33 for the outgoing line 20 of the link 10. Although shown as separate transmit and receive buffers, these functions may be combined. The controllers 28 or 29 are responsive to decoded command information in signals, cells or packets on the incoming lines 25 or 21 to activate the receivers 27 or 23 to start loading the respective buffers 30 or 31 with the incoming data, as well as to route the received data to be transmitted at one of the other ports. As seen in Figure 3, a cell 34 by which information may be conveyed from a node 13 to the node 11 is illustrated. This cell 34 is delineated by the underlying service, or by some element within the header (i.e., the header checksum). The cell begins with a header 35 which includes a virtual channel identifier 36, a control area 37 and a header checksum 38 used to verify the integrity of the header and of the framing. The payload field 39 is the major part of the cell 34. The controller 28 for a port to a link 14 is responsive to the virtual channel identifier 36 to control the routing of the incoming cell through the switching network 43 to attempt to pass the cell from one port to another in order to effect the virtual channel between source and destination. When there is a difference in bandwidth between the links 14 and the link 10, for example, the switching network 43 may include a multiplexer 44 to allow more than one link 14 to funnel into the link 10; likewise, a multiplexer 45 may allow simultaneous delivery of data from link 10 to more than one of the links 14. Similarly, the ports for links 14 may have multiplexers 46 and 47 so that data from or to multiple ports may be interleaved. Alternatively, the data may be interleaved by merely reading and writing between buffers 30-33 one word at a time via the switching circuit 43. In any event, a message frame is made up in transmit buffer 33, for example, by the controller 29, and this frame may contain interleaved packets or cells from many different terminals 17, going to many different terminals 18.

Referring to Figure 4, a message frame 50 used in the PTM technique is illustrated. This frame 50, employed for transmission on the link 10 in one example, is of fixed length 51 and is made up of a large number of slot cells 52. In an exemplary embodiment, the slot cells 52 each contain fifty-three octets (424-bits), and there are 2119 cells in a frame 50 of 6- millisecond length 51, transmitted at a rate of about 150-Mbps. A slot cell 52 contains a data field 53 of 48-octets and a header 54 of five octets; the header includes a channel identifying number associated with a particular transmission from a source to a destination node. The first two cells of the frame 50 are sync cells 55; these sync cells delimit each frame 50 which is sent during a loop control period. At least two sync cells 55 are sent at the beginning of each frame. Sync cells are identified by a specific header address, and each contains a pointer to the first slot in the control period (i.e., the first data cell 52 in the frame) which follows sync and slip cells. (The second sync cell contains a pointer value of one lower than the first sync cell.) At least one slip cell 56 follows the sync cells; a slip cell contains no information, other than a header address identifying it as a slip cell. These slip cells exist only to be added or discarded, as required, to synchronize the two sides of a loop (in node 11 and node 12, for example) when they are not running at identical speeds. Typically one slip cell 56 is sent after the sync cell 55, but a second will be added, or the one will be deleted, as required. The slot cells 52 are the ones assigned to carry slotted traffic, each having a valid virtual channel identifier in its header 54. Slot cells are carried with priority over free cells, and are allocated using control cells. A free cell 57 is an unallocated cell, and may be empty, in which case its header carries a virtual channel identifier for an empty cell, or may carry traffic for which no slot is assigned; this unallocated traffic is carried on a best-effort basis and may be discarded in favor of slot cells when allocated traffic appears. Finally, a control cell 58 is one that carries information (control signals, commands, etc.) between the two ends of the loop, e.g., from node 11 to node 12. A control cell 58 is identified by a specific virtual channel identifier in its header 59 (which may be locally assigned). Control cells carry messages that indicate that a given time slot within the basic control period of the frame 50 has been assigned to carry traffic on behalf of a given virtual channel, or has been freed. A protocol is defined for sending these messages in control cells, and the controller 29 generates these cells for sending via transmitter 22.

The frames 50 are timed by a separate clock 60 in each node 11 or 12. These clocks are stable crystal oscillators which maintain sufficient accuracy to synchronize the repetition of frames 50 at each end of a link 10. Since the time period 51 of a frame 50 is some multiple of 6-millisecond, this level of accuracy is well within that of currently-available crystal oscillators, or other methods of establishing a stable time reference. The bit rate, about 150-Mbps (or 600-Mbps, depending upon the network), is established by the oscillator 60, and the octet, cell and frame rates may be obtained by counting down from the bit rate clock, or from external synchronization sources, i.e., a network master clock.

Referring to Figure 5, an example of connection establishment is illustrated where a seven-hop connection (including two local loops) is shown between two of the terminals 17 and 18, labelled Y and Z in this example. Nodes 11, 11a and 12 are of the type shown in Figures 1 and 2, using the framing loops of Figure 4 in links 10 and 10a the link 10 is assumed to have an 11-millisecond one-way propagation delay, synchronized at four base periods or 24-ms. with 8192 slots 52 in a frame 50, while link 10a is assumed to have a 2-ms. one-way propagation delay, synchronized at one base period of 6-ms with 2048

slots. No assumption is made that bandwidth is symmetrical; traffic is to be passed from Y to Z, without regard for traffic from Z to Y (which is handled separately). When the connection is requested by terminal Y, the network controller first identifies the path from Y to Z, determined here to be a path

Y —> A —> B —> C —> D —> E —> Z.

Links A-B and D-E, like the local loops Y-A and E-Z, are local and can maintain an acceptably low loss rate by using conventional local credit management techniques. Links B-C and C-D (like link 10 of Figure 1) are longer and are phase-locked to one another and to the network master clock (or local synchronized clocks 60). Nodes B, C and D thus perform a phase-comparator function upon their plesiochronous transfer mode links so that a fixed mapping between slots in adjacent links is possible, using the frames 50 of Figure 4.

Once the path Y-to-Z is identified, the network controller (i.e., one of the controllers 29) allocates the appropriate bandwidth along each link. The unsynchronized links such as A-B and D-E need merely to have sufficient bandwidth available. The phase-locked links B-C and C-D have slots 52 allocated to the virtual channel Y-Z. Slots are assigned as far apart within the loop period 51 as possible, in order to minimize funneling effects. The number of slots allocated in each link is based on the bandwidth required. A link whose loop control period is greater than the basic control period of 6-millisecond is treated as having multiple instances of the basic control period. Thus in the example, if the channel Y Z requires a bandwidth of 256-kbps, then four slots are assigned in link C-D (which is operating at frame period of 6-ms.), each slot ideally following 512-slots after the previous one. The link B-C (which is operating at frame period of 24-ms.), however, requires sixteen slots assigned, again ideally 512-slots apart. If a link operated at a higher speed, the spacing between slots would remain uniform in time, and scaled in number of slots per frame. For example, if B-C were a 620 Mbps link, then it would have 32768 slots, and the sixteen slots in half-circuit Y Z would be spaced 2048 slots apart.

After the slots are allocated by the network controller, the network controller signals to Y by a control packet or cell that it is ready to accept traffic. Access node A grants credits to Y, and in turn forwards the cells it receives from Y on to B when it receives sufficient credits from B. B in turn inserts the cells into the appropriate time slots in the frame 50 currently being sent by B onto link B-C; C does the same in relaying them to D. Link D-E, however, is controlled by a simple credit mechanism so D buffers the cells until E has granted the required credits, at which time it forwards them to E, who in turn relays them asynchronously to Z.

Referring to Figure 6, a data stream sent on link 10a from a node 11a, for example, to the node 12 is a series of frames 50a, 50b, 50c, et seq., where each frame is of the format of Figure 4. The actual propagation delay 61 in the link 10a between nodes 11a and 12 is less than the length 51 of each of the frames 50a, 50b, etc. The receiver in the node 12 is synchronized by its clock 60 to receive the frame 50a beginning at time 62, and subsequent frames at times 63 spaced by the period 51, in a continuous sequence. If the clocks 60 drift, or the propagation delay drifts due to environmental factors, then there is some elasticity in the receive data buffers to account for such differences, while a drift of a cell length is accounted for adding or deleting slip cells 56. Drift of the order of magnitude of the bit rate (150- or 600-Mbps) or at the octet rate (18.75- or 75-M/sec.) is accounted for in the elasticity of the receive circuitry. The "phase- locking" referred to between transmitted and received frames in at slot level and frame level. Note that the slots or cells are phase locked between the various links in the network, and the frames are phase locked between a transmit-receive pair of nodes, but the frames 50 may be of different length in link 10 compared to frames in link 10a.

Significantly, and as described herein, a different allocation method is used for bursty traffic. This secondary procedure, used only for bursty (variable bit rate) traffic, employs a form of "fast circuit" switching, in which virtual channels are varied in size in order to handle variations in offered load. Access nodes or terminals such as node A of Figure 5 are expected to buffer traffic as it arrives, but when the buffer begins to fill, this originator node A may request a temporary increase in bandwidth sufficient to empty itself.

All virtual channels such as the channel Y Z have a residual bandwidth (BRes) which is available at all times; in the above example the residual bandwidth may be one slot per 512 slots, to be automatical- ly allocated by the network controller whenever a request is made, without any exchange of signals between nodes 11 and 12. Additional bandwidth for a given virtual channel is requested by means of a bandwidth request descriptor which is a message packet or cell 64 as seen in Figure 7, sent from the node 13 to the node 11, for example. This request descriptor message of course includes a field 65 to identify the source node and the destination node (or channel number) so that the path can be determined, and in addition has three elements, a BWext field 66, a BWquo field 67 and a duration field 68. The BWext field 66 specifies the requested bandwidth extension, and is a value representing the most that the network will grant for the duration of the descriptor. The network control facility determines what maximum bandwidth will ever be allotted to a terminal, depending upon the network configuration at the time, and sends this value to all nodes. The BWext field 66 is in the message 64 when it is sent by the originator to the

network. The BWquo field 67 is the bandwidth extension quota, which is the amount that the network actually grants. This value is initially set by the originating terminal 13 to be equal to BWext but may be lowered by any of the intermediate nodes 11, 11a, 12, etc., before being returned to the originator by the network. The duration field 68 is the amount of time that the descriptor should remain in effect, which is ideally expressed as a number of cells. While a time-based descriptor could be used, it would have to be extended by the network if the request BWext were reduced to a lower BWquo, since it would take longer to send the pending traffic.

Every node within a network such as that of Figure 5 also determines, at connection establishment, the total end-to-end transit delay across the network and its own relative position within the link (i.e., how many ms. from each end). When a message or bandwidth request descriptor or cell 64 is issued, this procedure is followed: (1.) The originator (e.g., node 13) sends a bandwidth descriptor 64 across the link towards the destination terminal Z. This cell 64 is identified as a user-to-network cell by its header 69. (2.) As the cell 64 travels towards the destination (using the transport mechanism of Figure 5, for example), each node 11, 11a 12, etc., determines how much of the requested bandwidth in BWext it can provide exactly one round-trip interval hence. If it cannot provide at least as much as the current BWquo in field 67 (which is being marked down as the descriptor 64 travels towards the destination), it puts a new value in BWquo field 67. No node may raise the value in BWquo field 67 as the cell 64 traverses the network. (3.) At the egress node (E in Figure 5), the descriptor 64 is returned along the same path by which it arrived. Each node (12, 11a, 11, etc.) on the return path notes the remaining (marked down) value of the BWquo field 67, but does not further change it; this value is stored in a table of all current traffic, identified by the channel number in the field 65, so that the controller 29 can check subsequent slotted cells for validity and also keep track of allocated capacity when making allocation for subsequent requests 64. (4.) When the descriptor 64 returns to its originator node 13, the bandwidth described in BWquo field 67 becomes available for immediate use by the terminal Y, for the time period of the duration field 68.

The concept of a virtual path instead of a virtual channel may be used to minimize node complexity. The total number of virtual channels between any two nodes such as Y and Z in a network is likely to frequently exceed one, e.g., when more than one message is pending. Some economization may take place by allocating bandwidth descriptors 64 to virtual paths instead of virtual channels. There can be only one virtual path between two nodes Y and Z. A virtual path, in this case, is a special form of channel that contains within itself multiple user virtual channels. The total number of virtual paths within a network is thus limited to the square of the number of nodes, regardless of the number of virtual channels requested by users. Access nodes can map each virtual channel into the appropriate virtual path, and intermediate nodes need only keep track of virtual paths (as well as any virtual channels that are locally terminated).

Referring to Figure 8, a chart of the allocated traffic in one link 10, for example, as a function of tine, shows that as the number of requested allocations from the remote nodes changes the allocated traffic level follows a line 71, rising and falling as the requests ebb and flow. A line 72 represents the limit imposed by the capacity of the link, determined by the physical construction, software, etc. During a peak in traffic, when the requested allocations from the remote terminals tend to exceed the limit 72 to follow the line 73, the controller imposes reduced (instead of requested) allocations on all remotes so the real traffic follows the line 74, below the limit 72, instead of the line 73. All network traffic is at a lower level than requested for a time period 75, until the line 76 rejoins the request curve 77 when all delayed requests have been made up. In this manner, network congestion during the peak period 75 is merely exhibited to the remote terminals as a slowing of the apparent response of the network, rather than as loss of data requiring retransmitting sequences of messages. Retransmission occurrences not only markedly reduce the apparent speed of the network from the terminals standpoint, but also reduce the real capacity of the network since traffic is transmitted more than once.

While this invention has been described with reference to specific embodiments, this description is not meant to be construed in a limiting sense. Various modifications of the disclosed embodiments, as well as other embodiments of the invention, will be apparent to persons skilled in the art upon reference to this description. It is therefore contemplated that the appended claims will cover any such modifications or embodiments as fall within the true scope of the invention.

Claims

1. A method of congestion avoidance in a communications network having multiple nodes, (11, 11a, 12) comprising the steps of:
   a) sending from a first of said nodes to a first intermediate one of said nodes a request 64) for an allocation of bandwidth for a transmission of a quantity of data from said first node to a second one of said nodes via said first intermediate node and a second intermediate node; characterized by:
   b) comparing (34) said request in said first

intermediate node with capacity at said first intermediate node to meet said request, and generating a modified request to reduce said allocation if necessary;

c) sending said modified request from said first intermediate node to said second intermediate node;

d) comparing said modified request in said second intermediate node with capacity at said second intermediate node to meet said modified request, and generating a second modified request to reduce said allocation if necessary;

e) sending said second modified request back to said first node; and

f) transmitting said quantity of data from said first node to said second node using the bandwidth specified in said second modified request.

2. A method according to claim 1 wherein said step of comparing said request in said first intermediate node includes comparing with capacity projected at a time about that required for one round trip between said first node and said second node.

3. A method according to claim 1 or 2 wherein said first node is allocated a standard bandwidth value and only sends said request if the bandwidth needed for transmitting said quantity of data exceeds said standard bandwidth value.

4. A method according to any of claims 1 to 3 wherein data is transmitted from said first intermediate node to said second intermediate node by an asynchronous transfer mode.

5. A communications network having multiple nodes (11, 11a, 12, 13, 15), comprising:

a) a transmitter (22) sending from a first of said nodes to a first intermediate one of said nodes a request for an allocation of bandwidth for a data transmission from said first node to a second one of said nodes via said first intermediate node and a second intermediate node;

b) means (34) for comparing said request in said first intermediate node with capacity at said first intermediate node to meet said request, and for generating a modified request to reduce said allocation if necessary, said modified request being sent from said first intermediate node to said second intermediate node;

c) means for comparing said modified request in said second intermediate node with capacity at said second intermediate node to meet said request, and for generating a second modified request to reduce said allocation if necessary, said second modified request being sent back to said first node;

d) said transmitter sending data from said first node to said second node using the bandwidth specified in said second modified request.

6. A network according to claim 5 wherein said means for comparing said request in said first intermediate node includes means for comparing with capacity projected at a time about that required for one round trip between said first node and said second node.

7. A network according to claim 5 or 6 wherein said first node is allocated a bandwidth value [a] equal to a standard allocation and only sends said request if the bandwidth exceeds said standard bandwidth value.

8. A network according to any of claims 5 to 7 including a transceiver in said first intermediate node which transmits data from said first intermediate node to said second intermediate node by an asynchronous transfer mode.

9. A network according to any of claims 5 to 8 wherein said transmitter sends data by a serial data link.

10. A method of transmitting a quantity of digital information from a node in a network to a destination, comprising the steps of:

a) transmitting from said first node a control cell containing a request for a bandwidth allocation needed for said quantity of information, a bandwidth quote equal to said bandwidth allocation, and a duration for the requested bandwidth;

b) receiving said control cell at an intermediate node in said network and marking down said bandwidth quote in said immediate node in response to the capacity of said intermediate node at a time when said quantity of information is to be transmitted, then transmitting said control cell toward said destination;

c) subsequently receiving at said first node a return of said control cell in which said bandwidth quote is modified downward in accordance with capacity of other nodes in said network;

d) transmitting from said first node to said destination said quantity of information at a bandwidth corresponding to said modified bandwidth quote.

11. A method according to claim 10 wherein said

duration is expressed in quantity of digital information.

12. A method according to claim 10 or 11 wherein said node is granted a residual allocation of bandwidth, and said control cell is transmitted only if said request exceeds said residual bandwidth.

13. A method according to any of claims 10 to 12 including the step of receiving said control cell at an intermediate node and marking down said bandwidth quote in said control cell at said intermediate node in response to the capacity of said intermediate node at a time when said quantity of information is to be transmitted.

14. A method according to any of claims 10 to 13 wherein said information is transmitted from said at least one node in said network to another node by an asynchronous transfer node.

15. Apparatus for transmitting a quantity of digital information from a node in a network to a destination, comprising:

a) means for transmitting from said node a control cell containing (1) a request for a bandwidth allocation needed for said quantity of information, (2) a bandwidth quote equal to said bandwidth allocation, and (3) a duration for the requested bandwidth;

b) means for receiving said control cell at an intermediate node in said network and modifying downward said bandwidth quote in said control cell in response to the capacity of said intermediate node at a time when said quantity of information is to be transmitted;

c) means for transmitting said control cell including said modified downward bandwidth quote from said intermediate node toward said destination;

d) means in said first node for subsequently receiving at said node a return of said control cell in which said bandwidth quote is modified downward in accordance with capacity of other nodes in said network;

e) and means in said node for transmitting from said node to said destination said quantity of information at a bandwidth corresponding to said modified downward bandwidth quote.

FIG.1

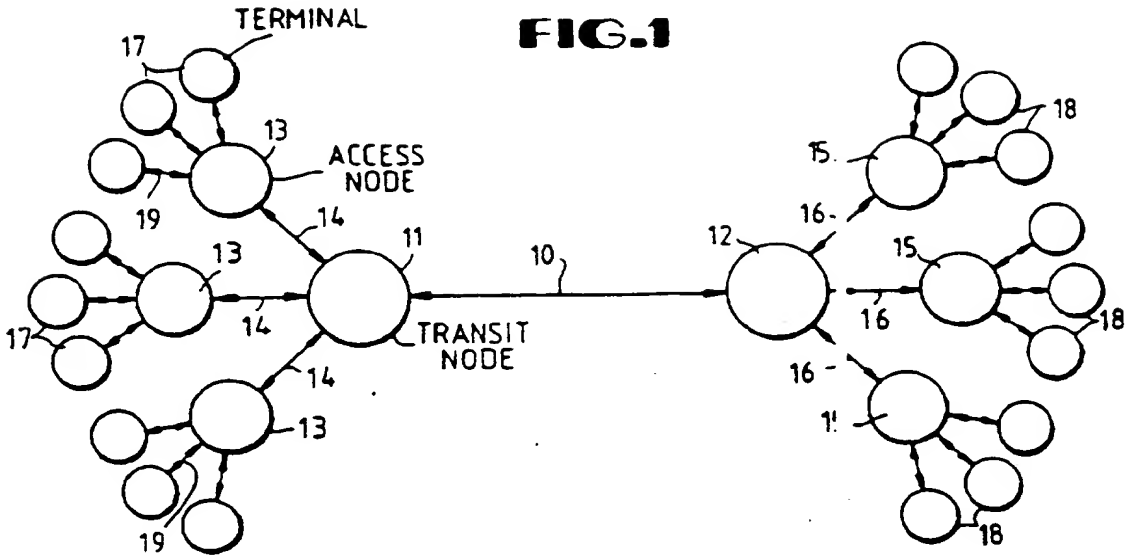
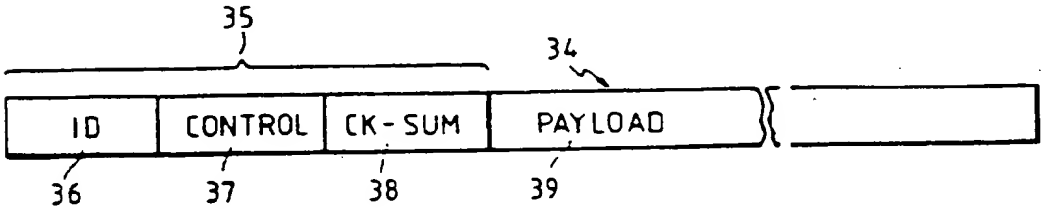


FIG.3



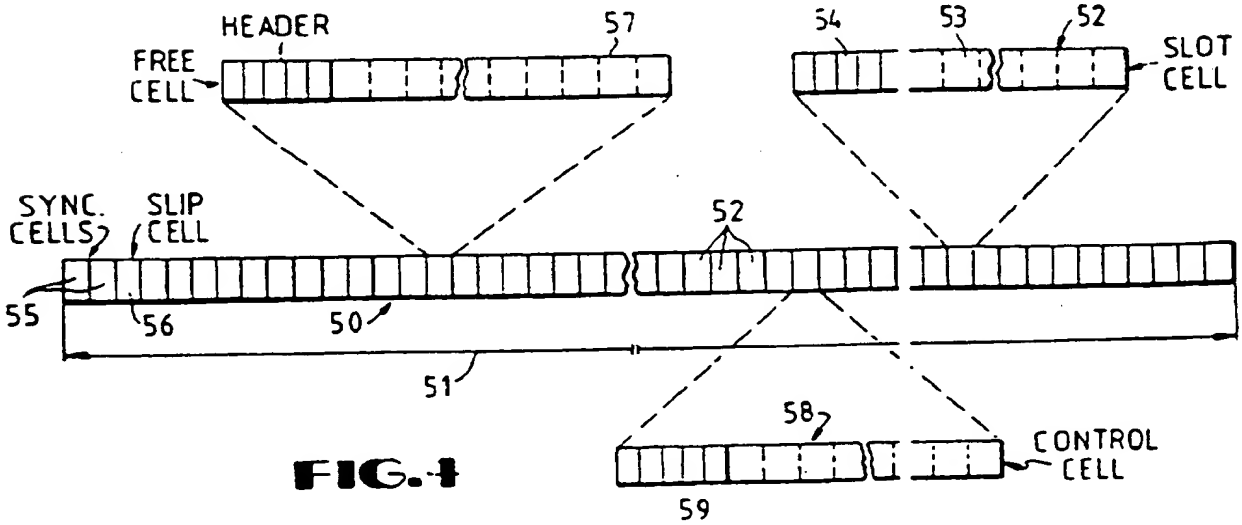


FIG.4

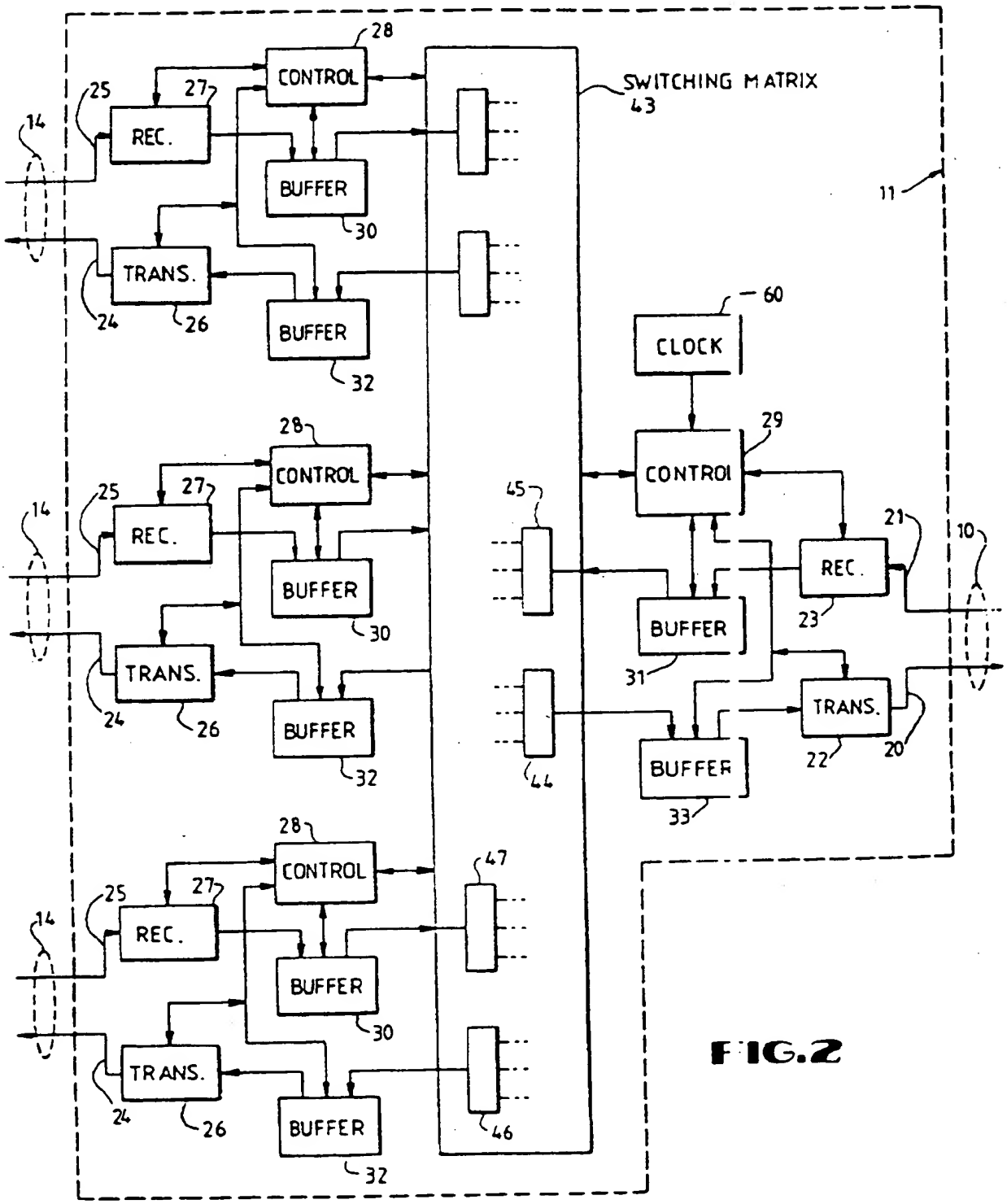


FIG.2

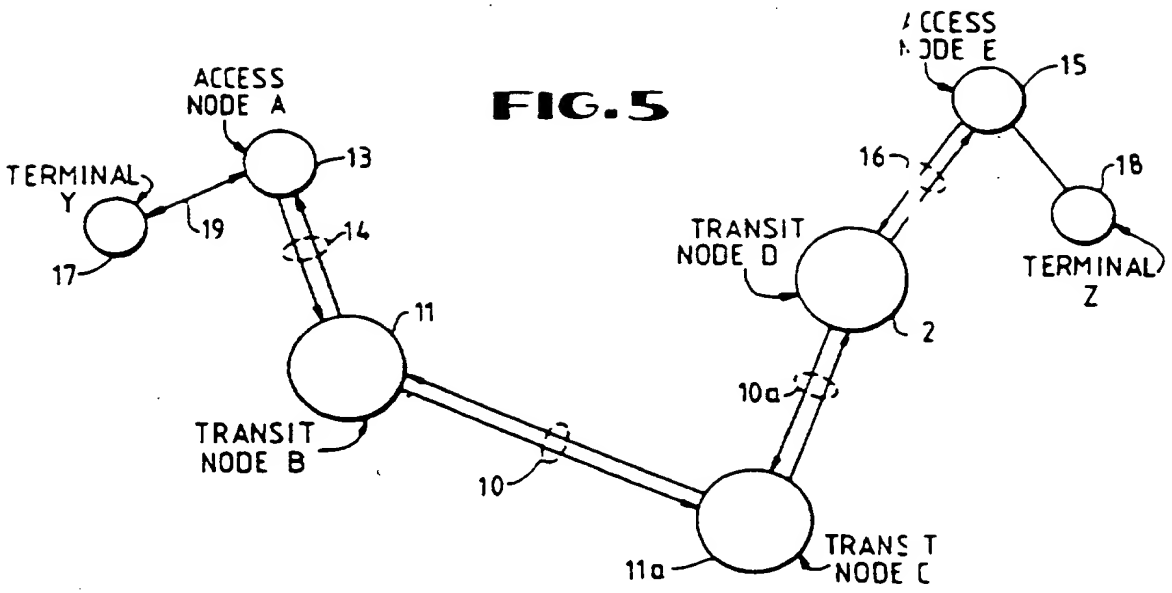


FIG.5

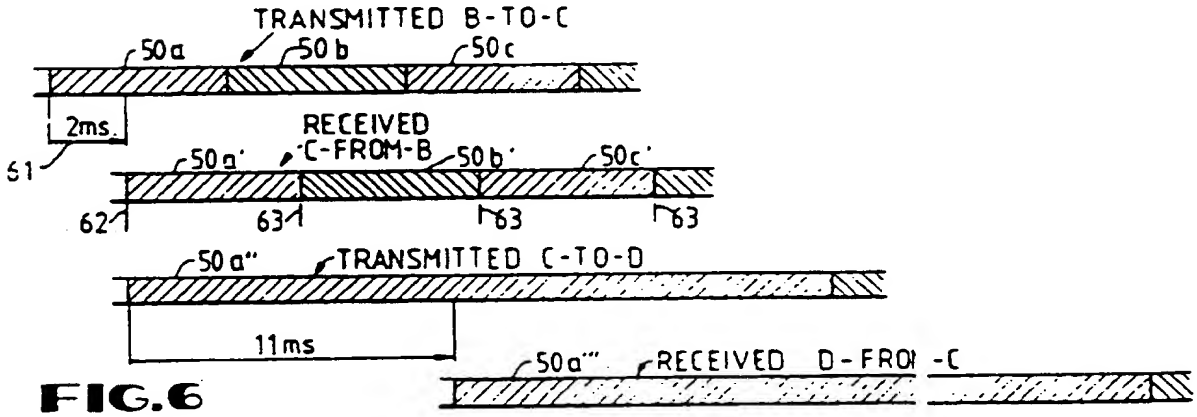


FIG.6

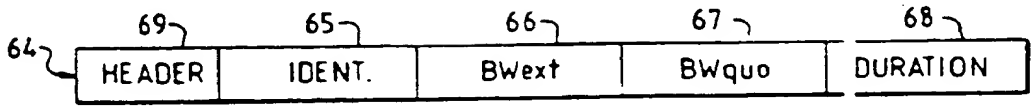


FIG.7

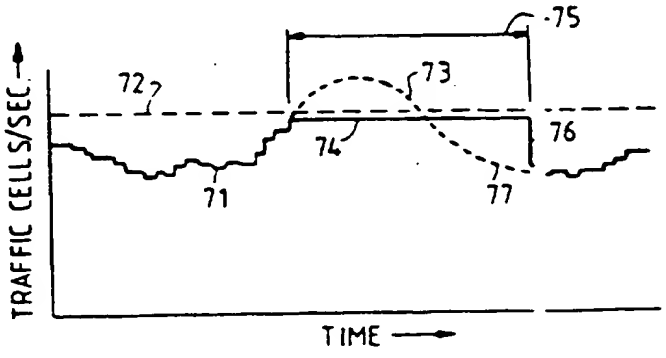


FIG.8

(19) Europäisches Patentamt
European Patent Office
Office européen des brevets

(13) Publication number : **0 453 355 A3**

(12) EUROPEAN PATENT APPLICATION

(21) Application number : **91400976.6**

(22) Date of filing : **11.04.91**

(51) Int. Cl.^{5} : **H04L 12/56** – V E

(30) Priority : **13.04.90 US 509605**

(43) Date of publication of application :
23.10.91 Bulletin 91/43

(84) Designated Contracting States :
DK FR GB IT

(88) Date of deferred publication of search report :
10.08.94 Bulletin 94/32

(71) Applicant : **DIGITAL EQUIPMENT CORPORATION**
146 Main Street
Maynard, Massachusetts 01745 (US)

(72) Inventor : **Goldstein, Fred Richard**
279 Appleton Street
Arlington, Massachusetts 02174 (US)
Inventor : **Callon, Ross**
5 Hilda Road
Bedford, Massachusetts 01730 (US)

(74) Representative : **Mongrédien, André et al**
c/o SOCIETE DE PROTECTION DES INVENTIONS
25, rue de Ponthieu
F-75008 Paris (FR)

(54) Congestion avoidance in high-speed network carrying bursty traffic.

(57) A data communication network subject to bursty traffic employs a bandwidth allocation scheme to avoid congestion. When a source node has a burst of traffic to send, it first sends a bandwidth request message through the network from source to destination. At each intermediate node, this bandwidth request is examined and the node determines how much of the requested traffic level it will be able to support at a time in the future of one round-trip interval hence, and this node either grants the request or marks down the request to a level that it can support, then passes it on. When the request reaches the destination, it is returned along the same path to the source, and the source then employs the marked-down allocation to select the rate used to send the burst of data. The allocation for this source node remains in effect for a limited time, depending upon the amount of data to be sent in the burst, then returns to a "residual" level.

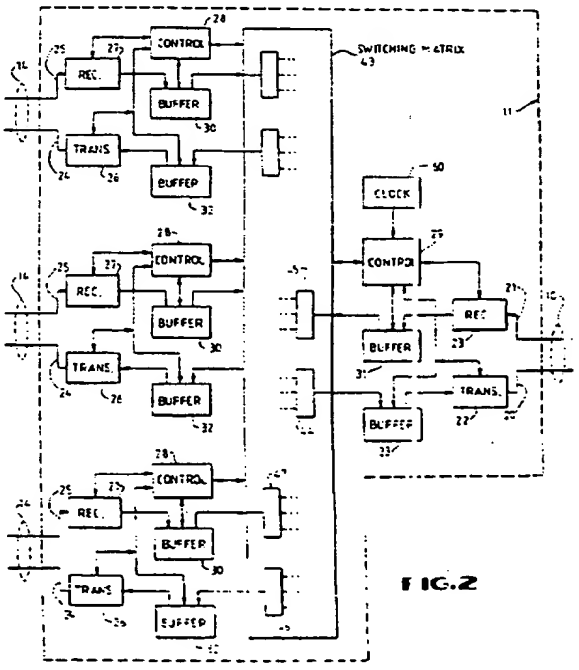


FIG.2

EP 0 453 355 A3

European Patent Office

EUROPEAN SEARCH REPORT

Application Number
EP 91 40 0976

DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document with indication, where appropriate, of relevant passages	Relevant to claim	CLASSIFICATION OF THE APPLICATION (Int.Cl.5)
X	IBM TECHNICAL DISCLOSURE BULLETIN. vol. 24, no. 4 , September 1981 , NEW YORK US pages 2044 - 2046 BHARATH-KUMAR ET AL. 'bottleneck flow control'	1,5	H04L12/56
Y		4,8,9	
A		10,15	
Y	EP-A-0 293 314 (ETAT FRANCAIS) 30 November 1988 * abstract *	4,8,9	
A	EP-A-0 249 035 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 16 December 1987 * column 3, line 22 - column 4, line 8 *	1,5,10, 15	
A	US-A-4 538 147 (GROW) 27 August 1985 * column 4, line 47 - line 64 *	1,5,10, 15	
			TECHNICAL FIELDS SEARCHED (Int.Cl.5)
			H04L

The present search report has been drawn up for all claims

Place of search	Date of completion of the search	Examiner
THE HAGUE	8 April 1994	Goossens, A

CATEGORY OF CITED DOCUMENTS

X : particularly relevant if taken alone
Y : particularly relevant if combined with another document of the same category
A : technological background
O : non-written disclosure
P : intermediate document

T : theory or principle underlying the invention
E : earlier patent document, but published on, or after the filing date
D : document cited in the application
L : document cited for other reasons

& : member of the same patent family, corresponding document

EPO FORM 1503 03.82 (P04C01)